Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.6, No.6, 2015
www.iiste.org

A Survey Paper on Sequence Pattern Mining with Incremental Approach

Rutva Patel
G.E.C,Modasa
rutva88@gmail.com

Prof. J.S. Dhobi
G.E.C,Modasa
jsdhobi@yahoo.com

Abstract:
Sequential pattern mining finds frequently occurring patterns ordered by time. The problem was first introduced by Agrawal and Srikant [1]. An example of a sequential pattern is "A customer who purchased a new Ford Explorer two years ago, is likely to respond favourably to a trade-in option now". Let X be the clause "purchased a new Ford Explorer" and Y be the clause "responds favourably to a trade-in". Then notice that the pattern XY above, is different from pattern YX which states that "A customer who responded favourably to a trade-in two years ago, will purchase a Ford Explorer now". The order in which X and Y appear is important, and hence XY and YX are mined as two separate patterns.Sequential pattern mining is widely applicable since many types of data have a time component to them. For example, it can be used in the medical domain to help determine a correct diagnosis from the sequence of symptoms experienced; over customer data to help target repeat customers; and with web-log data to better structure a company's website for easier access to the most popular links[2].

1. INTRODUCTION

An itemset is a non-empty set of items. A sequence is an ordered list of itemsets. Without loss of generality, we assume that the set of items is mapped to a set of contiguous integers. We denote an itemset i by $(i_1, i_2, i_3, \ldots i_m)$ where i_j is an item1. We denote a sequence s by $<s_1, s_2, s_3, \ldots s_n>$ where s_j is an itemset.

A sequence $<a_1, a_2, a_3, \ldots a_n>$ is contained in another sequence $<b_1, b_2, b_3, \ldots b_m>$ if there exists integers $k_1 < k_2 < \ldots < k_n$ such that $a_1 \subseteq b_{k1}, a_2 \subseteq b_{k2}, \ldots a_n \subseteq b_{kn}$. For example, the sequence <(3) (4 5) (8)> is contained in <(7) (3 8) (9) (4 5 6) (8)>, since (3) ⊆ (3 8), (4 5) ⊆ (4 5 6) and (8) ⊆ (8). However, the sequence <(3) (5)> is not contained in <(3 5)> and vice versa. The former represents items 3 and 5 being bought one after the other, while the latter represent items 3 and 5 being bought together.

A record supports a sequence s if s is contained in it. The support count is incremented only once per record. The support for a sequence is defined as the fraction of the whole data set that contains this sequence. If this support ≤ min_sup, then the sequence is frequent.
The four algorithms are surveyed one by one.

1.1 GSP

It is an Apriori based algorithm for sequential pattern mining[3]. The difference is that GSP inserts some constraints into the mining process, i.e., time constraints, and relaxes the definition of transaction. Moreover, it takes the taxonomies into account. For time constraints, maximum gap and minimal gap are defined to specified the gap between any two adjacent transactions in the sequence. If the distance between two transactions is not in the range between the maximum gap and the minimal gap, then the two transactions can not be taken as two consecutive transactions in a sequence.

1.2 ISM

The ISM algorithm, proposed by [4], is actually an extension of SPADE, which aims at considering the update by means of the negative border and a rewriting of the database. The first step of ISM aims at pruning the sequences that become infrequent from the set of frequent sequences after the update. One scan of the database is enough to update the lattice as well as the negative border. The second step aims at taking into account the new frequent sequences one by one, in order to make the information browse the lattice using the SPADE generating process.

1.3 FREESPAN

FreeSpan [6] was developed to substantially reduce the expensive candidate generation and testing of Apriori, while maintaining its basic heuristic. In general, FreeSpan uses frequent items to recursively project the sequence database into projected databases while growing subsequence fragments in each projected database. Each

projection partitions the database and confines further testing to progressively smaller and more manageable units.

A big advantage of FreeSpan over GSP is that it successfully confines pattern generation to progressively smaller projected databases. This not only reduces the number of candidates to be checked at any one time, but also limits candidate generation to only those items and known large-k patterns guaranteed to form at least one large-k+1 pattern (even then the number of candidates may be substantial). However, FreeSpan also has some non-trivial costs:

the growth of a subsequence is explored at any split point in a candidate sequence resulting in several possible new subsequences. For example, a frequent subsequence $\{cd\}^1$ and item b, could grow into any or all of the following subsequences if present: <bcd>, <(bc)d>, <b(cd)>, <bdc>, <(bd)c>, <cbd>, <c(bd)>, <dbc>, <d(bc)>, <cdb>, <(cd)b>, <dcb>. Even then, this is not the full range of possible subsequences if repeating items occur.

1.4 PREFIXSPAN

PrefixSpan [2] utilizes the method of database projection to make the database for next pass much smaller and consequently make the algorithm more speedy. The au- thors claimed that in PrefixSpan there is no need for candidates generation. It recursively projects the database by already found short length patterns. This pattern growth idea is similar to that in Apriori heuristic.

For example, with the frequent 1-sequence as the prefix, the projected database is shown in Table-1. The projected databases only contain the suffix of these sequences, by scanning the projected database all the length-2 sequential patterns that have the parent length-1 sequential patterns as prefix can be generated. Then the projected database is partitioned again by those length-2 sequential patterns.

Customer ID	Customer Sequence
1	<ac(bc)d(abc)ad>
2	<b(cd)ac(bd)>
3	<d(bc)(ac)(cd)>

Table-1

Large Itemsets	Projected Database
a	<c(bc)d(abc)ad> <c(bd)> <(_c)(cd)>
b	<(_c)d(abc)ad> <(cd)ac(bd)> <(_c)(ac)(cd)>
c	<c(bc)d(abc)ad> <(_d)ac(bd)> <(ac)(cd)>
d	<(abc)ad> <ac(bd)> <(bc)(ac)(cd)>

<a>	0			
<b>	(3 2 1)	0		
<c>	(3 3 2)	(2 3 2)	0	
<d>	(3 3 0)	(3 3 1)	(3 3 2)	0
	<a>	<b>	<c>	<d>

prefixSpan Mining

The main cost of PrefixSpan is the projected database scanning process. In order to improve the performance a bi-level projection method that uses the triangle S-Matrix is introduced. The main problem of PrefixSpan, is the time consuming on scanning the projected database, which may be very large if the original dataset is huge.

2. CONCLUSION

Mining sequential patterns from the large transactional database is a very crucial task. There are many approaches that have been discussed; nearly all of the previous studies were using GSP approach and PrefixSpan approach for extracting the sequential patterns, which have scope for improvement. Thus the goal of this research was to find a scheme for pulling the sequential patterns out of the transactional data sets considering the time consumption.

References

[1] R. Agrawal; R. Srikant; , "Mining sequential patterns," *In Proceedings of Inter- national Conference on Data Engineering*, pp. 3–14, 1995

[2] Pei, J.; Han, J.; Pinto, H.; Chen, Q.; Dayal, U.; Hsu; , "PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth," *Proceedings of 2001 International Conference on Data Engineering*, pp. 215–224, 2001

[3] R. Srikant; R. Agrawal; , "Mining sequential patterns: Generalizations and performance improvements," *In Proceedings of International Conference on Ex- tending Database Technology*, pp. 3–17, 1996

[4] S. Parthasarathy; M. Zaki; M. Ogihara; S. Dwarkadas; , "Incremental and interactive sequence mining," In Proceedings of the 8th International Conference on Information and Knowledge Management (CIKM' 99), pp. 251–258, 1999

[5] J. Ayres; J. Gehrke; T. Yiu; J. Flannick; , "Sequential pattern mining using a bitmap representation," *In Proceedings of ACM SIGKDD International Confer ence on Knowledge Discovery and Data Mining,* pp. 429–435, 2002

[6] J. Han; J. Pei; B. Mortazavi-Asl; Q. Chen; U. Dayal; M-C. Hsu; , "FreeSpan: Frequent Pattern-Projected Sequential Pattern Mining," *In Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'00),* pp. 355-359, 2000

[8] Jiawei Han and Micheline Kamber, Data MiningConcepts and Techniques - Third Edition, ELSEVIER Morgan Kaufman Publisher, July 6, 2011

[9] D. N. Goswami, Anshu Chaturvedi, C. S. Raghuvanshi,"Frequent Pattern Mining Using Record FilterApproach",International Journal of Computer Science,Vol. 7, Issue 4, No 7, July 2010, pp 38-43

[10] Anjan K Koundinya,Srinath N K,K A K Sharma, Kiran Kumar, Madhu M N and Kiran U Shanbag, "Map/Reduce Design And Implementation Of Apriorialgorithm For Handling Voluminous Data-Sets", ACIJ, Vol.3, No.6, November 2012, pp 29-39

[11] Lamine M. Aouad, Nhien-An Le-Khac and Tahar M. Kechadi, "Distributed Frequent Itemsets Mining in Heterogeneous Platforms", Journal of Engineering, Computing and Architecture, Vol. 1, Issue 2, 2007

[12] Bagrudeen Bazeer Ahamed and Shanmugasundaram Hariharan, "A Survey On Distributed Data Mining Process Via Grid", International Journal of Database Theory and Application, Vol. 4, No. 3, September 2011, pp 77-90

[13] Florent Masseglia; Pascal Poncelet; Maguelonne Teisseire;, "Incremental mining of sequential patterns in large databases," *Data & Knowledge Engineering,* pp. 97-121, 2003

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

Academic conference: http://www.iiste.org/conference/upcoming-conferences-call-for-paper/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar